Lec 21 MATH 343

Non-example = corona

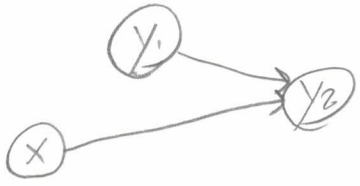
X = disease 1 (binary) = Heart Attacks

Y_1 = disease 2 (binary) } two downstream responses

Y_2 = hospitalization (binary)

Here's the DAG:

(E)



We are interested in testing if disease 1 is related to disease 2.

There's no causal arrow from X to Y_1 $\Rightarrow$ no relationship.

If we run the logistic regression $\ln\left(\frac{\hat{p}(Y=1)}{\hat{p}(Y=0)}\right) = b_0 + b_1 x \Rightarrow b_1 \approx 0$ and insignificant.

However, if we run the logistic regression only on the subset $\mathbb{1}[y_2 = 1]$

$\ln\left(\frac{\hat{p}(Y=1)}{\hat{p}(Y=0)}\right) = b_0 + b_1 x \Longrightarrow b_1 > 0$ and significant

Berkson's Paradox (1946)

What happens? Since both diseases are necessary to put someone in the hospital, if you're in the hospital, the diseases are suddenly related "collider bias"

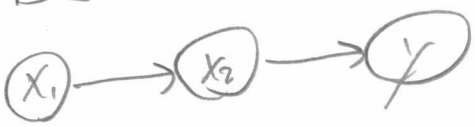
"collider bias"

$\Rightarrow$ This illustrates the pitfall of not using simple random samples. In order to estimate both a causal effect, you must sample from the whole population!

$\Rightarrow$ Is the correlation between X and Y_1 real for subjects who $y_2 = 1$?

Yes! But it is not causal. It is still useful for prediction!

(F)

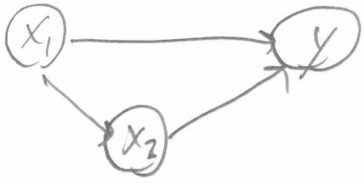


~~X~~: lung cancer
X_1: smoking
X_2: DNA damage

Is X_1 causal? Yes. $\hat{y} = b_0 + b_1 x_1$ will b_1 be unbiased? yes!

What about $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$. Will b_1 be unbiased? No!
The path is blocked!

(G) This is now the opposite of Simpson's Paradox



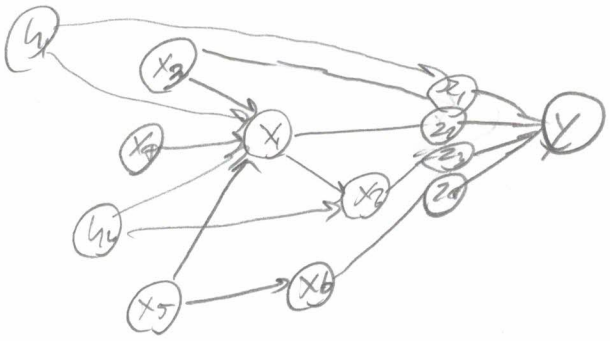
Is x_1 causal? Yes. $\hat{y} = b_0 + b_1 x_1$ will be unbiased? Yes!

What about $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$. b_1 is now biased. The path is partially blocked!

Let's sum up our scenarios: Y = response, X_1 = variable of interest
$X_2, \ldots, X_p$ = other variable

What is a <u>controlled</u> experiment / test / trial / study?

Consider the real but unknown DAG:

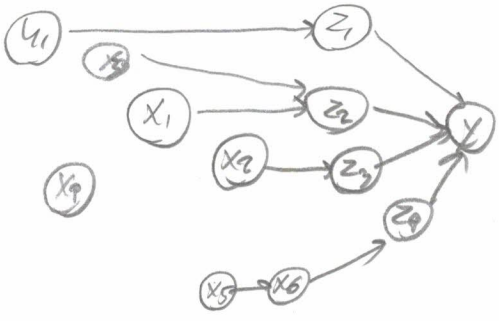


If we run the regression of y on all observed covariates ⇒ wrong

" " " " " only x ⇒ wrong

" " " " " you any subset of observed covariates ⇒ wrong

If we now do "surgery" on reality and delete all inputs into X_1, we have



This deletion of inputs now clears the path for a $\hat{y} = b_0 + b_1 x$ being unbiased for the causal effect of x_1 on y.

How to delete all inputs? Control / manipulate values of x_1 yourself.

If it's not manipulated it is called an "observational study" as the data is merely sampled and observed "in the wild".

Is this always possible? x_1 = Smoking, y = lung cancer (gets)

Can you force people to smoke? Can you force people not to smoke? No

Is this always ethical? x_1 = experimental (taking) and dangerous drug

y = Cancer (gets)

No!

If possible, is it practical and affordable?

x_1 = perfect (has) lab-grown hygienic diet at \$1000/meal, y = survival

But, if possible, affordable and ethical, controlled trials can provide causal estimates. It is not a panacea; all problems from 341 are still present: type I / type II errors / power.

Now... the details of experiments. Consider x_1 to be a binary variable and call it w so that $\vec{w} \in \{0,1\}^n$

$$\mathcal{D} = \left(\underbrace{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{w} & \vec{x}_{\cdot 1} & \cdots & \vec{x}_{\cdot p} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{X}, \vec{y} \right)$$

$\vec{w}$ is said to be a "binary manipulation", "binary treatment".

The subgroups $\{i : w_i = 0\}$ and $\{i : w_i = 1\}$ are called the "two arms" of the study. Typically, the $w=0$ is a "placebo" or "no treatment" or "business as usual" and the $w=1$ is the "new treatment" in which case if $\vec{y} = b_0 + b_1 w$, b_1 is the causal estimate of the new treatment over no treatment on y. Called "pill-placebo" controlled trial

If $w=0$ denotes treatment A and $w=1$ denotes treatment B, then b_T is the causal estimate of the difference of B over A. This is sometimes called a "comparative study" or an "A-B test".

The vector $\vec{w}$ is called the "assignment" or "allocation" since you are assigning / forcing / manipulating subjects to receive w_i. How do we generate $\vec{w}$'s? There are 2^n possibilities! The means of generating $\vec{w}$'s is called the "design". What design is "best"? Difficult problem! My whole research program of the past 10yr focuses on it!

To measure "best" we need to focus on specific settings and make assumptions.